GOSSIP ALGORITHMS AND THEIR VARIANTS

Vivek Borkar

IIT Bombay

December 12, 2014

Outline

Classical ('*vanilla*') gossip

Random gossip

Optimal gossip

Nonlinear gossip

‘Gossip’ algorithm

$$x_i(n+1) = \sum_{j=1}^{d} p(j|i)x_j(n), \; n \geq 0.$$

$P = [[p(j|i)]]_{1 \leq i,j \leq d}$ irreducible stochastic matrix with unique stationary distribution $\pi \Longrightarrow x(n) \to \pi^T x(0)\mathbf{1}$.

Research focus on rate of convergence: Design a ‘good’ P ((doubly) stochastic, low |second eigenvalue|, $\cdots$) (Boyd, Shah, Ghosh, $\cdots$)

Ref: ‘*Gossip Algorithms*’, D. Shah, NOW Publishers, 2009.

Often a component of a ‘larger’ scheme:

$$x_i(n+1) = (1-a)x_i(n) + a\sum_{j=1}^{d} p(j|i)x_j(n) + \cdots, \; n \geq 0.$$

Examples: Distributed computation, Synchronization, ‘Flocking’, Coordination of mobile agents

The objective often is ‘consensus’.

The DeGroot model

Models opinion formation in society.

$$x_i(n+1) = (1-a)x_i(n) + a\sum_{j=1}^{d} p(j|i)x_j(n), \ n \geq 0.$$

New opinion a convex combination of own previous opinion and opinions of neighbors/peers/friends.

Convergence $\Longrightarrow$ asymptotic agreement.

What about **random** gossip?

$$x_i(n+1) = (1-a)x_i(n) + a x_{\xi_{n+1}(i)}(n),$$

where $\xi_n(i)$ IID $\approx p(\cdot|i)$.

Convergence?

Yes!!

And **consensus:** $x(n) \to c\mathbf{1}$, but c may not be $\pi^T x(0)$!

Analysis based on re-writing the iteration as

$$x_i(n+1) = (1-a)x_i(n) + a\sum_{j=1}^{d} p(j|i)x_j(n) + aM_j(n+1),$$

where $\{M(n)\}$ is a martingale difference sequence. This is a '*constant step-size stochastic approximation*'.

Fact: Standard 'intuition' would suggest asymptotically a random walk along the degenerate direction $c\mathbf{1}, c \in \mathcal{R}$, but we still get convergence because 'noise' $\{M(n)\}$ is also killed asymptotically at a fast enough rate.

But what if we want the actual average $\pi^T x(0)$?

Alternative scheme based on the ‘Poisson equation’:

for $f(i) = x(0)$,

$$V(i) = f(i) - \beta + \sum_j p(j|i)V(j), \ 1 \leq j \leq d. \qquad (1)$$

Solution $(V(\cdot), \beta)$ satisfies: β unique, $= \pi^T f$, V unique up to additive scalar.

Can solve (1) by the ‘relative value iteration’

$$V^{n+1}(i) = f(i) - V^n(i_0) + \sum_j p(j|i)V^n(j), \ n \geq 0.$$

The ‘offset’ $V^n(i_0)$ stabilizes the iteration, other choices are possible (e.g., $\frac{1}{d}\sum_k V^n(k)$).

'*Reinforcement learning*': stochastic approximation version of RVI – for a simulated chain $\{X_n\} \approx p(\cdot|\cdot)$.

$$\begin{aligned} V^{n+1}(i) &= (1 - a(n)I\{X_n = i\})V^n(i) + \\ &\quad a(n)I\{X_n = i\}(f(i) - V^n(i_0) + V^n(X_{n+1})). \end{aligned}$$

Then $V^n(i_0) \to \beta$ a.s.

(**Not** fully decentralized: needs $V^n(i_0)$ to be broadcast. Can replace it by $\frac{1}{d}\sum_k V^n(k)$ which can be calculated in a distributed manner by another gossip on a faster time scale.)

'Multiplicative' analog of the previous case: for $f(i) > 0$, choose $V^0(i) > 0 \ \forall \ i$ and do:

$$V^{n+1}(i) = \frac{f(i)\sum_j p(j|i)V^n(j)}{V^n(i_0)}, \ n \geq 0.$$

More generally, for irreducible nonnegative $Q = [[q(i,j)]]$, set

$$f(i) = \sum_k q(i,k), \ p(j|i) = \frac{q(i,j)}{f(i)}.$$

Then $V^n(i_0) \to$ the Perron-Frobenius eigenvalue of Q, $V^n \to$ the corresponding eigenvector.
('power' method)

Applications : ranking, risk-sensitive control

‘Learning’ version: for $V^0(\cdot) > 0$,

$$V^{n+1}(i) = (1 - a(n)I\{X_n = i\})V^n(i) + a(n)I\{X_n = i\}\left(\frac{f(i)V^n(X_{n+1})}{V^n(i_0)}\right).$$

Numerically better even when the eigenvalue is known!

(The first term on RHS scales slower than the second.)

Similar evolution occurs in models of emergent networks (Jain - Krishna)

OPTIMAL GOSSIP

Gossip for opinion manipulation (e.g., advertising):

$P_1 :=$ submatrix of P corresponding to $n-m$ rows and corresponding columns,

$P_2 :=$ submatrix of P corresponding to the same $n-m$ rows and remaining m columns.

These m columns correspond to nodes whose 'opinion' is frozen at x^*. Then we have (in $\mathcal{R}^{n-m}$):

$$x(n+1) = x(n) + a(n)\left[P_1 x(n) + P_2 x^* \mathbf{1}\right].$$

Assume P_1 strictly sub-stochastic, irreducible. Then:

$x(n) \to x^* \mathbf{1}$ exponentially at rate $\lambda :=$ the Perron-Frobenius eigenvalue of P_1.

$\Longrightarrow$ consensus on a pre-specified value.

Objective: Minimize λ over all subsets of cardinality m (i.e., find the m most important nodes for information dissemination)

Hard combinatorial problem, even the nonlinear programming relaxation is highly non-convex and the projected gradient scheme with multi-start does not do too well.

$\Longrightarrow$ Use '**engineer's licence**'.

For $\tau :=$ the first passage time to frozen nodes,

$\lambda = -\lim_{t\uparrow\infty} \frac{1}{t} \log P(\tau > t)$ and $E[\tau] = \sum_{t=0}^{\infty} P(\tau \geq t)$.

$\Longrightarrow$ Use $E[\tau]$ as a surrogate cost.

This is *monotone and supermodular* $\Longrightarrow$ greedy scheme is $\left(1 - \frac{1}{e}\right)$-optimal (Nemhauser-Wolsey-Fisher)

Important observation: best m nodes $\neq$ top m nodes according to individual merit!

What about controlling the transition probabilities?

Consider controlling the nonlinear o.d.e.

$$\dot{x}(t) = \alpha(P_1^{u(t)} - I)x(t) + \alpha P_2^{u(t)}(x^* \mathbf{1}) + (1-\alpha)F(x(t))$$

with ‘cost’

$$E\left[\int_0^\infty e^{-\beta t} \sum_i |x_i(t) - x^*|^2 dt\right].$$

Here $P_\cdot^u = [[p(j|i,u)]]$.

Can write down the corresponding Hamilton-Jacobi-Bellman equation and verification theorem.

$\Longrightarrow$ Optimal

$$u_i^*(t) \in \text{Argmax}\left(\sum_{j=1}^{n-m} p(j|i,\cdot)x_j^*(t) + x^* \sum_{j=n-m+1}^{n} p(j|i,\cdot)\right)$$

for $x < x^*$, and,

$$u_i^*(t) \in \text{Argmin}\left(\sum_{j=1}^{n-m} p(j|i,\cdot)x_j^*(t) + x^* \sum_{j=n-m+1}^{n} p(j|i,\cdot)\right)$$

for $x > x^*$.

($\Longrightarrow$ greatest ‘push’ towards x^*.)

NONLINEAR GOSSIP

STOCHASTIC APPROXIMATION

Consider the Robbins-Monro scheme in $\mathcal{R}^d$:

$$x(n+1) = x(n) + a(n)[h(x(n)) + M(n+1)].$$

Here:

- $h : \mathcal{R}^d \mapsto \mathcal{R}^d$ Lipschitz,

- $\{M(n)\}$ a martingale difference sequence w.r.t. $\mathcal{F}_n := \sigma\left(x(m), M(m), m \leq n\right), n \geq 0$, i.e.,

$$E\left[M(n+1)|\mathcal{F}_n\right] = 0.$$

Also, there exists $K \in (0, \infty)$ such that

$$E\left[\|M(n+1)\|^2 | \mathcal{F}_n\right] \leq K\left(1 + \|x(n)\|^2\right).$$

- Step-sizes $a(n) > 0$ satisfy:

$$\sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty.$$

‘ODE Approach’ (Derevitskii-Fradkov-Ljung)

View the iteration as a noisy discretization of the ODE

$$\dot{x}(t) = h(x(t)), \; t \geq 0.$$

This is well posed under our hypotheses.

Definition: A set A is invariant if

$$x(0) \in A \Longrightarrow x(t) \in A \; \forall \; t \in \mathcal{R}.$$

Definition (continued):

A is *Internally Chain Transitive* if given any $x, y \in A$, and $\epsilon > 0, T > 0$, we can find $n \geq 1$, and

$$x = x_0,\ x_1,\ \cdots,\ x_{n-1},\ x_n = y \in A$$

such that for $0 \leq i < n$, the trajectory $x^i(t), t \geq 0$, of

$$\dot{x}^i(t) = h(x^i(t)),\ x^i(0) = x_i,$$

satisfies $\|x^i(t) - x^{i+1}\| < \epsilon$ for some $t \geq T$.

Benaim's theorem:

If $\sup_n \|x(n)\| < \infty$ a.s., then $x(n) \to$ a compact

connected nonempty internally chain transitive

invariant set of the ODE, a.s.

THE TSITSIKLIS MODEL

- ‘Agents’/processors placed at the nodes of an irreducible directed graph $\mathcal{G}$ with node set $\mathcal{V}$ with $|\mathcal{V}| := N$ and edge set $\mathcal{E}$. $\mathcal{N}(i) := \{i\text{'s neighbors}\}$.

- For $i \in \mathcal{V}$ and $P = [[p(j|i)]]$ stochastic, $\mathcal{G}$-compatible,

$$x_i(n+1) = \sum_j p(j|i)x_j(n) + a(n)[h(x_i(n)) + M_i(n+1)].$$

- At each instant, every node takes,
  - a weighted average of its neigbhbors' values (**'gossip' component**), and,
  - adds a correction based on its own computation (**'learning' component**).

- Delays, asynchrony, etc. (shall worry about it later).

Similar models in synchronization, flocking/coordination, ...

Objective: **CONSENSUS**

Nonlinear gossip I: quasi-linear case

For each $i \in \mathcal{V}$, consider the d-dimensional iteration

$$x_i(n+1) = \sum_{j \in \mathcal{N}(i)} p_{x(n)}(j|i) x_j(n) + a(n) \left[h_i(x_i(n)) + M_i(n+1) \right].$$

Here, P_x is an irreducible stochastic matrix where $x \mapsto P_x$ is Lipschitz, with $(\min)_j^+ p_x(j|i) \geq \Delta > 0$.

For a fully distributed algorithm, the ith row of $P_{x(n)}$ should depend only on $x_j(n)$, $j \in \mathcal{N}(i) \cup \{i\}$, but we use $x(n)$ without loss of generality.

Let $\pi_x :=$ the unique stationary distribution under P_x.

CONSENSUS:

if $\sup_{i,n} \|x_i(n)\| < \infty$ a.s., then

$$\|x_i(n) - x_j(n)\| \to 0 \text{ a.s.}$$

(Not surprising, standard arguments work.)

MAIN RESULT ($d = 1$):

Let $\mathcal{A} := \{c\mathbf{1} : c \in \mathcal{R}\}$. Let $x(n) = [x_1(n), \cdots, x_N(n)]^T$.

If $\sup_{i,n} \|x_i(n)\| < \infty$ a.s., then almost surely,
$x(n) \to \mathcal{A}_0 :=$ an internally chain transitive invariant set of N-fold copy of the ODE

$$\dot{y}(t) = \sum_k \pi_{y\mathbf{1}}(k) h_k(y(t)), \ t \geq 0,$$

contained in $\mathcal{A}$.

General case: Define

$$\mathcal{A} := \{x = [(x^1)^T : \cdots : (x^N)^T]^T \in \mathcal{R}^{d \times N} : \\ x^i = [x^i_1, \cdots, x^i_d]^T, 1 \leq i \leq N;\ x^i_k = x^j_k\ \forall\ i, j\}.$$

Consider

$$\dot{y}(t) = \sum_{i=0}^{N} \pi_{\psi(y(t))}(i) h_i(y(t)).$$

where $\psi(y) := [y^T : y^T : \cdots : y^T]^T$ for $y \in \mathcal{R}^d$.

Then $\mathcal{A}$ is invariant under this dynamics.

Theorem $\sup_n \|x_n\| < \infty$ a.s. $\Longrightarrow x(n) \stackrel{n\uparrow\infty}{\rightarrow}$ a compact connected non-empty internally chain transitive invariant set $\mathcal{A}_0 \subset \mathcal{A}$ of the N-fold product of the above dynamics, a.s.

(That is, dynamics in $\mathcal{R}^N$ wherein each component satisfies the above o.d.e.)

Stronger results possible for special cases
(e.g., convergence for $d = 1$!)

Example: Consider $h_i = -\nabla f \ \forall i$. Let $|\mathcal{N}(i)| = M \ \forall i$ and for a prescribed $T > 0$ ('temperature')

$$
\begin{aligned}
p_x(j|i) &= \frac{1}{M} e^{-\frac{(f(x_j)-f(x_i))^+}{T}}, \ j \in \mathcal{N}(i), \\
&= 0, \qquad\qquad j \notin \mathcal{N}(i), j \neq i, \\
&= 1 - \sum_{k \in \mathcal{N}(i)} p_x(k|i), \quad j = i.
\end{aligned}
$$

Then

$$
\pi_x = \frac{e^{-\frac{f(x_i)}{T}}}{\sum_j e^{-\frac{f(x_j)}{T}}}.
$$

This puts more weight on low values of f (spatial annealing).

Can think of this scheme as a '*leaderless swarm*' by analogy with *Particle Swarm Optimization*, wherein each particle uses information from self, neighbors, and the 'best so far', i.e., a leader. Here the last piece is 'emergent' from a distributed gossip.

Another example: Dependence of P_x on x due to mobility.

A *'stability test'*: Define

$$g(x) := \sum_i \pi_x(i) h_i(x),$$

$$g_c(x) := \frac{g(cx)}{c} \text{ for } c > 0,$$

$$g_\infty(x) := \lim_{c \uparrow \infty} g_c(x),$$

assumed to exist. Then g_c, g_∞ are Lipschitz.

Consider the ODE ('scaling limit')

$$\dot{x}_\infty(t) = g_\infty(x_\infty(t)),\ t \geq 0.$$

If this has the origin as the unique asymptotically stable equilibrium, then $\sup_n \|x(n)\| < \infty$ a.s.

Intuition: Iterates large in absolute value track this o.d.e. after scaling, hence exhibit stabilizing drift.

Nonlinear gossip II: fully nonlinear case

$$x_i(n+1) = f_i(x(n)) + a(n)\left[h_i(x_i(n)) + M_i(n+1)\right], \; i \in \mathcal{V}.$$

- $f := [f_1, \cdots, f_N]^T : (\mathcal{R}^d)^N \mapsto (\mathcal{R}^d)^N$ is continuous, and,
- $P(x) = \lim_{n\uparrow\infty} f^{(n)}(x)$ ($:= f \circ f \circ \cdots \circ f$, n times) exists, with the limit being uniform on compacts. (Then $P(P(x)) = P(f(x)) = f(P(x)) = P(x) \in C := \{x : P(x) = x\}$.)

Assumptions:

1. P is Frechet differentiable with its Frechet derivative $\bar{P}_x(\cdot)$ continuous in x.

2. $\bar{P}_{f(\cdot)}h(\cdot)$ is Lipschitz. (Ideally, should be 'local', but we ignore this issue.)

3. $E\left[\|M(n+1)\|^4|\mathcal{F}_n\right] \leq F(x(n))$ for some continuous F.

Assume $\sup_n \|x(n)\| < \infty$ a.s.

Consider the ODE

$$\dot{x}(t) = \bar{P}_{x(t)}(h(x(t))).$$

MAIN RESULT: $x(n) \to$ a compact connected nonempty internally chain transitive invariant set of the above ODE contained in C, a.s.

Example: $P :=$ a projection to a convex set, $x(n+1) = f(x(n))$ an iterative scheme for calculating the projection.

In this case, we get a projected version of the distributed stochastic approximation scheme.

$\Longrightarrow$ Need distributed scheme for computing projections on, e.g., intersection of convex sets.

COMING SOON: A distributed version of the Boyle-Dykstra-Han scheme*

*joint work with Soham Phade

Some standard issues in distributed computation:

1. Interprocessor delays

2. Asynchrony: not all updates at the same time

3. Updates may be on ‘local clock’

Replace

$$x_i(n+1) = f_i(x(n)) + a(n)[\ \cdots\ \ \cdots\ ]$$

by

$$x_i(n+1) =$$

$$(1 - b(\nu(i,n))I\{i \in B(n)\})x_i(n)\ +\ b(\nu(i,n))I\{i \in B(n)\}$$

$$\times\ f_i(x_1(n - \tau_{1i}(n)), \cdots, x_N(n - \tau_{Ni}(n)))\ +$$

$$a(\nu(i,n))I\{i \in B(n)\}[h_i(x_1(n - \tau_{1i}(n)), \cdots) + M_i(n+1)],$$

with $\sum_n b(n) < \infty,\ \sum_n b(n)^2 < \infty,\ a(n) = o(b(n))$.

Here,

- $B(n) :=$ { nodes 'active' at time n},

- $\nu(i,n) :=$ # updates by i till time n. Need:

$$\liminf_{n\uparrow\infty} \frac{\nu(i,n)}{n} > 0 \text{ a.s.}$$

  This ensures that all processors update comparably often.

- $\tau_{ji}(n) :=$ the delay with which j's output was received by i at time n,

  i.e., at time n, i has access to $x_j(n - \tau_{ji}(n))$, but not $x_j(m), m > n - \tau_{ji}(n)$.

- Additional conditions on stepsizes.

  Among them: if $\tau(t), t \geq 0$, denotes the time scaling ('algorithmic' or 'ODE' time scale) given by

$$\tau(n) := \sum_{m=0}^{n-1} b(m), \; n \geq 0,$$

  with linear interpolation on each $[n, n+1]$, then

$$\lim_{n \uparrow \infty} \frac{\tau(\alpha n)}{\tau(n)} \to 1 \; \forall \; \alpha \in (0, 1).$$

  For example, $b(n) = \frac{1}{n} \Longrightarrow \tau(t) \approx \log t$ will do.

Under above modifications, earlier results hold:

1. Bounded delays ‘squeezed out’ (i.e., they lead to asymptotically negligible error) due to time scaling (more generally, conditional moment conditions suffice)

2. Asynchrony / local clocks compensated for by the choice of stepsize (get back the original limiting ODE modulo time-scaling)

References

1. VB, R. Makhijani, R. Sundaresan, **Asynchronous gossip for averaging and spectral ranking**, *IEEE J. Selected Topics in Signal Processing* 8(4), 2014.

2. VB, A. Karnik, U. Jayakrishnan Nair, S. Nalli, **Manufacturing consent**, to appear in *IEEE Transactions on Automatic Control*.

3. A. S. Mathkar, VB, **Nonlinear gossip**, *submitted*.